Toxicity in ChatGPT: Analyzing Persona-assigned Language Models

Year: 2023 | Citations: 442 | Authors: A. Deshpande, Vishvak Murahari, Tanmay Rajpurohit, A. Kalyan, Karthik Narasimhan

---

Abstract

Large language models (LLMs) have shown incredible capabilities and transcended the natural language processing (NLP) community, with adoption throughout many services like healthcare, therapy, education, and customer service. Since users include people with critical information needs like students or patients engaging with chatbots, the safety of these systems is of prime importance. Therefore, a clear understanding of the capabilities and limitations of LLMs is necessary. To this end, we systematically evaluate toxicity in over half a million generations of ChatGPT, a popular dialogue-based LLM. We find that setting the system parameter of ChatGPT by assigning it a persona, say that of the boxer Muhammad Ali, significantly increases the toxicity of generations. Depending on the persona assigned to ChatGPT, its toxicity can increase up to 6x, with outputs engaging in incorrect stereotypes, harmful dialogue, and hurtful opinions. This may be potentially defamatory to the persona and harmful to an unsuspecting user. Furthermore, we find concerning patterns where specific entities (e.g., certain races) are targeted more than others (3x more) irrespective of the assigned persona, that reflect inherent discriminatory biases in the model. We hope that our findings inspire the broader AI community to rethink the efficacy of current safety guardrails and develop better techniques that lead to robust, safe, and trustworthy AI systems.